Open Data, Social Media and other User Generated Text: some topics for a discussion session

Ineke Schuurman, Leen Sevens, Vincent Vandeghinste

May 2018, the new personal data protection –privacy! – rules (GDPR) will become valid in the EU and all its member states. On the other hand, making your research data available as Open Data is becoming more and more important, for example when applying for a grant (NWO, FWO, EU, ...). Not making them 'Open' is to be justified very well. What does this mean for linguistic research? Especially when dealing with children, elderly people, people with an intellectual disability, migrants, ... Or with social media? And what when someone decides that his or her contribution to a corpus can no longer be used? Thus: How are GDPR and Open Data to be reconciled? How are you dealing with such issues? Any tips, tricks, ideas, solutions? Or more topics that should be taken care of? *** One of the pros for all of us of working with Open Data, is that data will be reusable. For example, we ourselves are working on data written by people with an intellectual disability (ID), data with lots of specific errors (spelling, grammar), or just consisting of pictos. These data can only be used by us. That's a pity, as it is difficult to get such data, especially digital, for example from social media. Some issues we would like to discuss with you in this session, esp. when you are working with user-generated data (social media, web forums, online reviews, ...). * in general, how would you ask permission from everybody involved? Depending on the type of resource, people may be using nicknames, ... * how would you deal with users with some communicative issues, in general or wrt the language in question (apply text normalization, translation, ...? If so: on request or in general?) * and what with informed consent in such, from the user and/or a carer, guardian, ... And what about the ethical aspects? * how would you act when people recall their permission? * in such a case: what about older versions of your data already made available to other reseachers? * which metadata would you be make available to other researchers? All, or ...? Would you allow them to see the real names, etc? We presume that all people appearing in the data collection will be anonymized or pseudonymized (unless the names of public figures are involved.) !! We will try and get answers to issues arising during this session after CLIN 2018 (for example during our ISI-NLP 2 workshop at LREC).